Choosing Adequate Speech Presence Probability Method for Enhanced Multichannel Background Noise Reduction Algorithms

Iván Darío Arévalo Gutiérrez

Universidad Nacional de Colombia
Facultad de Ingeniería, Departamento de Ingeniería Eléctrica y Electrónica.
Bogotá, Colombia
2017

Choosing Adequate Speech Presence Probability Method for Enhanced Multichannel Background Noise Reduction Algorithms

Iván Darío Arévalo Gutiérrez

Trabajo de grado presentado como requisito parcial para optar al título de:
Pregrado en Ingeniería Electrónica

Director:
Jan Bacca Rodríguez Ph.D.

Universidad Nacional de Colombia
Facultad de Ingeniería, Departamento de Ingeniería Eléctrica y Electrónica.
Bogotá, Colombia
2017

"The voice of beauty speaks softly; it creeps only into the most fully awakened souls"
- Friedrich Nietzsche

Acknowledgment

I want to give special thanks to all the pedagogic staff at the Universidad Nacional de Colombia, who guided me to achieve my dreams.

Also, I would like to thank all the team of the Universidad de Colombia Radio (UN Radio), who gave their time and knowledge to help me recording an appropriated data base for this project.

Finally I would also like to thank all my family and my friends, who always believed in me and have given me all the support I need to continue.

Abstract

There are several voice communication systems that are used nowadays which are capable of maintain voice calls between two users in real time. Telephones are widely used all around the world in an unlimited kind of situations. All of these situations expose the microphone (or microphones) of the phones to different and unpredictable noises, as street noise, sea noise, rain noise, wind noise, unwanted voices, car motors, etc. As microphones capture all the sounds around it, including the wanted voice and the unwanted noises, it is necessary to implement digital real time filters capable of attenuate as much as possible all the surrounding noises.

It exists a large quantity of noise reduction methods that have been used in the calling algorithms of phones. Even if these methods have had, in general, a good performance, there is still a research being done in this area in order to improve the current results. Because of this, the multichannel methods were created (using multiple microphones) as well as new algorithms that pretend to have a better noise reduction than the single channel methods. Most of these methods require a speech presence probability (SPP) method to achieve the noise reduction.

The following document presents a research about different SPP methods as well as a comparison between these. This includes an explanation on how theses algorithm work, a Matlab implementation using real voice and noise recordings and objective tests of the filter.

Key Words: Noise suppression, Speech processing, Speech enhancement, Adaptive signal processing, Active filters, Digital filters, Noise, Low-frequency noise, Noise cancellation, Signal to noise ratio.

Contents

1 Introduction

Nowadays the use of the mobile phones is generalized all around the world and it has become a device which is used in all contexts of life. Phone calls are done in all different kinds of environments, exposing the microphone (or microphones) to a large kind of noises. It is required that the noises captured by the microphone system is not transmitted in the phone call, because of this, digital real time filters are used. This complete process is often called voice enhancement.

Several works in the frequency domain have used arrays with multiple microphones to improve the previous singles channel approaches. One of these research is the one found in [18], there it is proposed a multichannel speech enhancement method based on the Wiener filter which is incorporated to a single channel filter that is used as reference. This algorithm has proved to have better performance in noisy situations over the conventional single channel methods.

As most of the speech enhancement methods, the one based on the multichannel Wiener filter, needs one Voice Activity Detector (VAD) to perform the calculation of the noise and voice correlation matrices. In [18] they propose a variation which uses a Speech Presence Probability (SPP) method to achieve this. The SPP contains the information of how likely it is that there is presence of voice in a frequency bin.

As several VAD and SPP methods have been developed for different algorithms, the goal of this project is to do a research of different of this methods, implement them on Matlab for a multichannel Wiener filter, do the right tests and show the dependency of this filter from the SPP or VAD block.

This document will be divided in 5 parts: problem and its background, where the basis of the multichannel Wiener filter algorithm are shown; Literature Review, where some previous work on the area will be found; Implementation, where it will be shown how the filter was implemented and testes; results, with comparison of the different variables and conclusions.

2 State of Art

Several methods have been developed in order to have a better speech enhancement. The state of art will be divided in two groups of algorithms: single channel and multiple channel. It is important to note that this algorithms can be mixed to obtain different kinds of noise reduction methods.

2.1 Single Channel Algorithms

2.1.1 Voiced Activity Detection (VAD)

Many methods of speech enhancement use the VAD method to make a difference between silence and speech periods. This periods are always differentiated by each of the STFT frames processed. The voiced sounds are periodic and usually have more energy that unvoiced sounds, also, unvoiced sounds are more noise-like and have more energy than silence segments.
The VAD can be used in many implementations as speech recognition, voice compression, noise estimation and suppression, and echo cancellation. For this, some methods have been developed. Probably the simplest VAD is Energy Level Detection [4], in which the initial noise spectrum, mean, and variance are calculated assuming the initial frames are only noise. After this, the thresholds are calculated for speech and noise decisions and all statistics are gradually updated when a noise frame is detected. The process diagram can be seen in figure **2-1**.
Some other methods can be found in [4], [14],[13], [5],[10]. Some of them are listed:

- Zero crossing rates: these can be calculated for each frame and compared with a threshold. The zero crossing rate of noise is assumed to be considerably larger than the one of speech. This assumption works well at high SNR values, but has problems at low SNR and in the presence of periodic noise.
- Periodicity: The detector uses a least-squares periodicity estimator, LSPE, on the input signal and triggers when a significant amount of periodicity is found.
- Speech Presence Probability (SPP): it is shown [5] that it is possible to create a limited maximum likelihood estimator, then use it for the estimation of the a priori signal-to-noise ratio (SNR), and this resulting noise power estimate can be updated when the a posteriori SNR is below a certain threshold and, finally, this threshold can be seen as a voice activity detector.

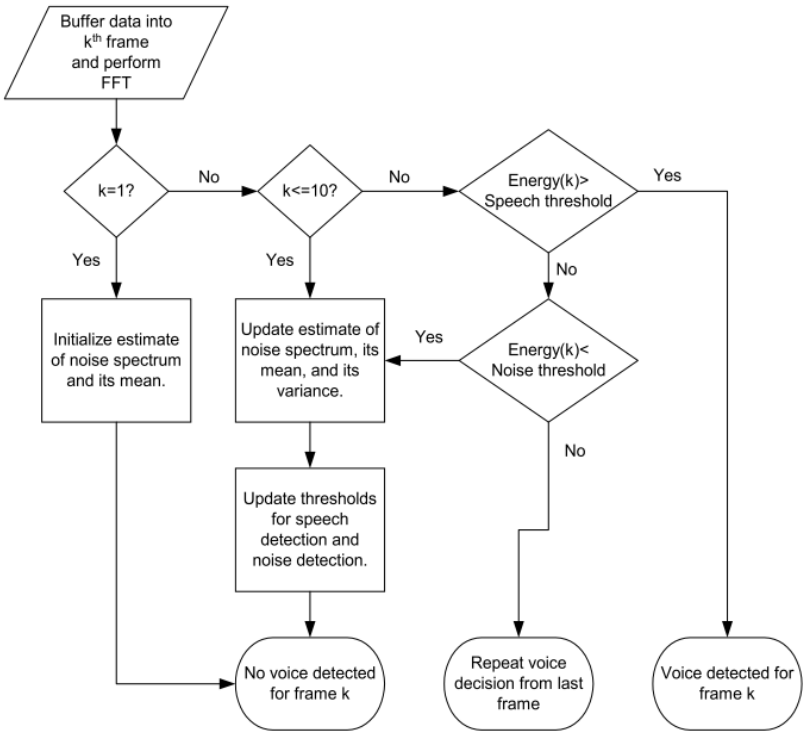

Figure 2-1: VAD with energy detection[4].

2.1.2 Spectral subtraction

This algorithm uses the short-term spectral magnitude of the noisy speech and estimate or reference of the noise signal. Most of the single channel spectral subtraction methods use a VAD to determine if there is silence or not in order to get an accurate noise estimate and the noise is assumed to be short term stationary so that noise from silence frames can be used to remove noise from voiced frames as shown in figure **2-2**.

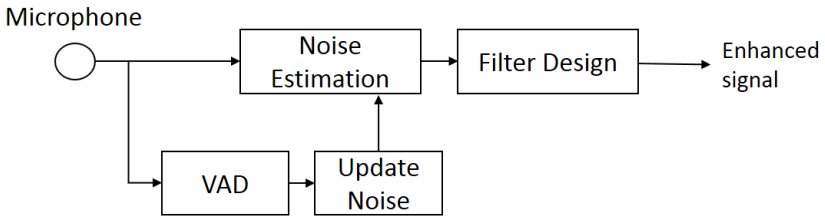

Figure 2-2: Basic Spectral Subtraction.

Many articles show different ways of using the spectral subtraction as [3], [6], [16], [1].
The spectral subtraction always uses a VAD. This is used to know the noise power only in the silence periods and this information is used to erase the noise from the speech as we can see in the figure **2-3**.

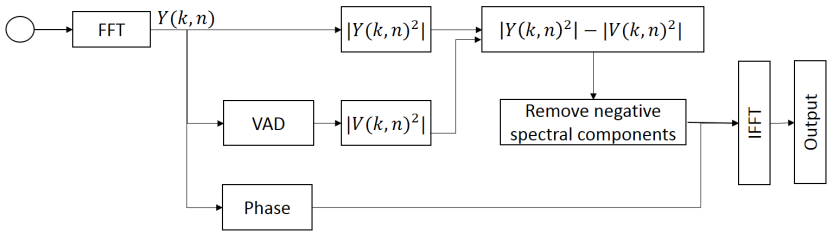

Figure 2-3: Single Channel Spectral Subtraction.

2.1.3 Talker isolation

It is necessary to isolate the desired speech from other speech sources and noise. In this case, separating the sources is crucial since the desired and undesired speech have similar spectra and comparable amplitudes.
It is used the frequency and amplitude continuity to track the desired talker. Binaural spatial cues are used to discriminate pitch frequencies that are too close to resolve spectrally. In [8] they show an example using a combination of techniques for advanced pitch tracking and talker isolation, it is used frequency and amplitude continuity to track the desired talker.

2.2 Multiple Channel Algorithms

2.2.1 Sources separation

The separation of sources is an approach in which the source signals are estimated from the mixed signals observed in each channel. This technique is used in sound systems in order to erase noise or to erase the cross-talking in communications. This system can be modeled as seen in figure **2-4** (with 2 sources).

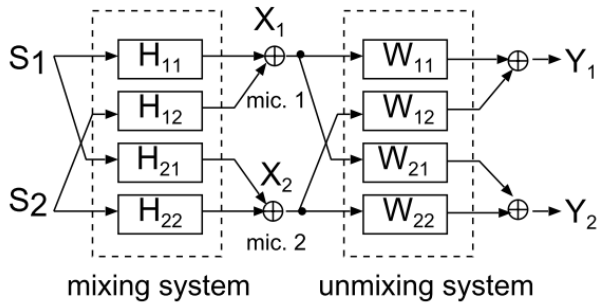

Figure 2-4: Separation of sources system configuration.

Many methods have been developed to achieve a proper separation of sources. In [16] it is shown a method in which, as some other methods like [2] and [9], every filter of the unmixing

system (W_{ij}) is found by an estimation of the mixing system (H_{ij}).

2.2.2 Beamforming

Beamforming is the term used for steering an array of sensors (microphones) to have unity gain in the direction of the desired source and attenuate the signals with origin in other directions. Source localization can be important to improve the effect of beamforming [4]. This can be used as a noise reduction method due to it's capacity of localization, that means that the noise (which comes from different directions than the voice) will be attenuated.
The technique of beamforming is used to capture multiple sound inputs $s(t)$ in the farfield and, after filtering and addition obtain a reference signal Z [4].
One technique called the "delay and sum beamformer" , which is probably the simplest, consist in apply a phase delay to the input signals to steer the main lobe directivity to an specific direction.

2.2.3 Coherence and Speech Presence Probability

In [10] is proposed one noise estimation method based in the speech presence probability value showed in [5] and the coherence function, which behaves as a measure of spectral similarity between signals. The coherence function, due to the noise characteristics of noise in diffuse field, has low values for noisy segments and a high value for speech segments.

2.2.4 Multi-Channel Wiener Filter

The Wiener filter is a filter used to produce an estimate of a desired or target, assuming known stationary signal and noise spectra, and additive noise. The goal of the this filter is to find a statistical estimate of an unknown signal using a reference signal as an input and filtering this signal to produce the estimate as an output. In [18] it is proposed a multichannel method which also uses one single channel method as reference. This algorithm will be described in detail later since it was chosen for implementation.

3 The Problem and its Background

As it was shown in the introduction, this work pretends to prove several VAD and SPP methods for the multichannel Wiener filter algorithm for speech enhancement. In this chapter will be explained how this filter works and why it is important to test the performance of the SPP block.

3.1 Signal Model

To maintain mathematical coherence during the document, the signal model will always be as it follows:
The microphone signals will be denoted as:

$$Y_m(k,n), l = 1, ..., M$$

Where k is the frequency bin index, n is the frame index and M is the number of microphones. So, the input signals are given by:

$$Y_m(k,n) = X_m(k,n) + V_m(k,n)$$

Where X_m and V_m are the target signal and the uncorrelated noise components. The goal is to remove the unwanted noise and preserve the target signal, this can be done by using a filter set $\mathbf{w}(k,n)$ to obtain:

$$Z(k,n) = \mathbf{w}^H(k,n)\mathbf{y}(k,n)$$

Where Z is the output signal and $\mathbf{y}(k,n)$ is a vector given as:

$$\mathbf{y}(k,n) = [Y_1(k,n), Y_2(k,n), ..., Y_M(k,n)]^T$$

$$\mathbf{y}(k,n) = \mathbf{x}(k,n) + \mathbf{v}(k,n)$$

The correlation matrices of noisy speech, clean speech and noise are defined as:

$$\mathbf{R}_y = E\{\mathbf{y}\mathbf{y}^H\} \in \mathbb{C}^{M\times M}$$

$$\mathbf{R}_x = E\{\mathbf{x}\mathbf{x}^H\} \in \mathbb{C}^{M\times M}$$

$$\mathbf{R}_v = E\{\mathbf{v}\mathbf{v}^H\} \in \mathbb{C}^{M\times M}$$

3.2 Multi-Channel Wiener Filter

Many of the real time implementations of the multichannel Wiener filter (MWF) have estimation problems caused mainly for using a voice activity detector (VAD), which may fail in adverse environments and the use of second order clean speech statistics, which usually causes overestimation errors. In this document it will be described the algorithm proposed in [18], first using the hard VAD and then using a SPP. Also, in order to have a better performance it will be chosen one of the input signals from the microphone array and will be filtered by a single channel algorithm and this new signal will be used as reference (the complete block diagram is shown in figure **3-1**). Finally it will be shown the objective tests that were applied to this algorithm in order to show a comparison with other methods.

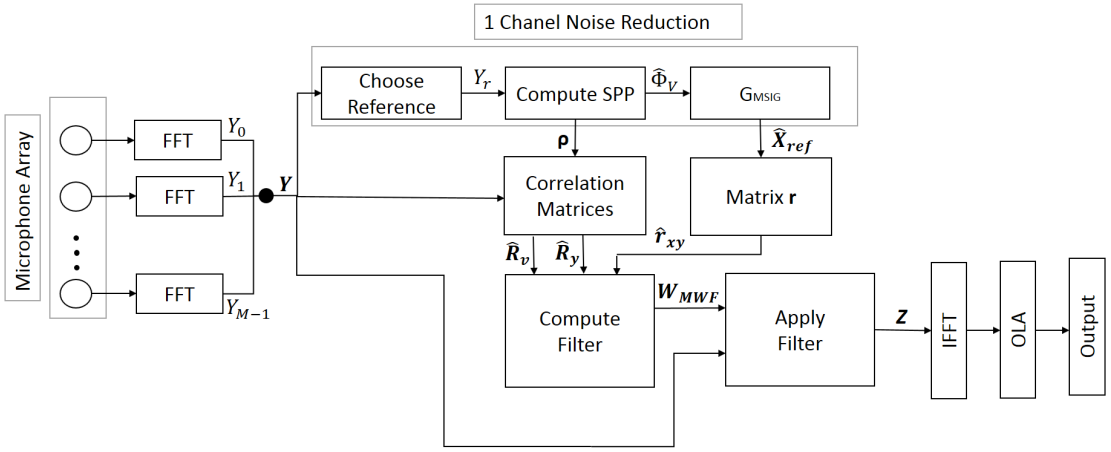

Figure 3-1: Block diagram for multichannel Wiener filter.

3.2.1 Filter Formulation

It is expected to estimate the speech signal based on an minimum mean square error criterion as:

$$\mathbf{w}_{MWF} = argmin_w E\{|X_{ref} - \mathbf{w}^H \mathbf{y}|^2\} \tag{3-1}$$

If it is taken the assumption of uncorrelation between speech and noise, it is possible to modify the criterion of the MWF as:

$$\mathbf{w}_{MWF} = argmin_w E\{|X_{ref} - \mathbf{w}^H \mathbf{x}|^2\} + \mu E\{|X_{ref} - \mathbf{w}^H \mathbf{v}|^2\}$$

A larger μ value here indicates more residual noise reduction at the expense of higher speech distortion. The solution of MWF_μ can then be obtained as:

$$\mathbf{w}_{MWF_\mu} = [\mathbf{R}_x + \mu\mathbf{R}_v]^{-1}\mathbf{R}_x\mathbf{e}_{ref} \tag{3-2}$$

Where $\mathbf{e}_{ref} = [0...010...0]^T$ is a M-element zero vector with the unity corresponds to the m^{th} element of the microphones.
In first place, the update of the correlation matrices can be done like:

$$\mathcal{H}_0 : \begin{cases} \hat{\mathbf{R}}_v[n] = (1-\alpha_v)\hat{\mathbf{R}}_v[n-1] + \alpha_v\mathbf{y}[n]\mathbf{y}^H[n] \\ \hat{\mathbf{R}}_y[n] = \hat{\mathbf{R}}_y[n-1] \end{cases} \tag{3-3}$$

$$\mathcal{H}_1 : \begin{cases} \hat{\mathbf{R}}_y[n] = (1-\alpha_y)\hat{\mathbf{R}}_y[n-1] + \alpha_y\mathbf{y}[n]\mathbf{y}^H[n] \\ \hat{\mathbf{R}}_v[n] = \hat{\mathbf{R}}_v[n-1] \end{cases} \tag{3-4}$$

Where $\mathcal{H}_0$ and $\mathcal{H}_1$ denote the periods of absence and presence of speech. This periods are determined by the use of a VAD algorithm. The choice of the smoothing factors α must be done carefully taking care if the degree of stationarity of speech and noise signals.
In equation (3-2) it is needed an estimation of $\mathbf{R}_x$ which could be obtained form $\mathbf{R}_y - \mathbf{R}_v$, but this estimation might have some errors caused by the complex valued matrices, which will lead to a bad estimation. This estimation could be improved by obtaining a pre-determined R_x estimate with a calibration sequence, or by implementing a mathematical model, these methods rely on the a priori information, which makes them less stable and difficult to use for final users.

3.2.2 Formulation of MWF_λ and Estimation of Correlation Matrices

To minimize the error from equation (3-1), another criterion to minimize the noise power is proposed. For this, one weighed sum can be find as:

$$\mathbf{w}_{MWF_\lambda} = argmin_w(1-\lambda)E\{|X_{ref} - \mathbf{w}^H\mathbf{y}|^2\} + \lambda(E\{|\mathbf{w}^H\mathbf{v}|^2\})$$

Where λ is a weighting value between 0 and 1, this solution is given by:

$$\mathbf{w}_{MWF_\lambda} = [(1-\lambda)\mathbf{R}_y + \mathbf{R}_v]^{-1}(1-\lambda)\mathbf{r}_{yx} \tag{3-5}$$

$$\lambda = 1 - \rho$$

where $\mathbf{r}_{yx} = E\{\mathbf{y}X^*_{ref}\}$. Using this consideration, it is avoided the estimation of $\mathbf{R}_x$ and the value of λ can be set as $1-\rho$ where ρ is the SPP of the signal. Also, it is proposed to avoid the decision by the VAD to estimate the correlation matrices and to use instead a modified SPP (ρ) with which the matrices can be estimated as:

$$\hat{\mathbf{R}}_v[n] = (1-\tilde{\alpha}_v[n])\hat{\mathbf{R}}_v[n-1] + \tilde{\alpha}_v[n]\mathbf{y}[n]\mathbf{y}^H[n] \tag{3-6}$$

$$\hat{\mathbf{R}}_y[n] = (1-\tilde{\alpha}_y[n])\hat{\mathbf{R}}_y[n-1] + \tilde{\alpha}_y[n]\mathbf{y}[n]\mathbf{y}^H[n] \tag{3-7}$$

Where $\tilde{\alpha}_v$ and $\tilde{\alpha}_y$ are given by:

$$\tilde{\alpha}_v = \alpha_v(1-\rho)$$

and

$$\tilde{\alpha}_y = \alpha_y\rho$$

Where α_v and α_y are the fixed smoothing factors for noise and noisy correlation matrices and ρ is the value of the speech presence probability.

Smoothing the filter

In the first implementation, as proposed in [18], the output signal tends to have musical noises. To solve this, it is proposed to do a smoothing process over the whole filter $\mathbf{W}_{MWF_\lambda}$ as:

$$\mathbf{W}_{MWF_\lambda}(k,n) = \alpha_w\mathbf{W}_{MWF_\lambda}(k,n-1) + (1-\alpha_w)[(1-\lambda)\mathbf{R}_y + \mathbf{R}_v]^{-1}(1-\lambda)\mathbf{r}_{yx} \tag{3-8}$$

Which leads to an important musical noise reduction.

Single-Channel Reference

From equation (3-5), we can see that this method needs the estimate of the speech reference X_{ref}, to achieve this, it is proposed to estimate X_{ref} by using a single-channel speech enhancement method and to use one microphone as a reference. With this, it is possible to define:

$$\tilde{\mathbf{r}}_{yx} = \mathbf{y}G(X^*_{ref} + V^*_r)$$

Where $X_{ref} = H_{ref}S$, in which H_{ref} is the acoustic transfer function of the target speech signal S, at the reference channel. In this equation G is a spectral weighting gain function, which involves the computation of the *a posteriori* and *a priori* SNR estimates. Also, to avoid musical noise, it is proposed to update recursively $\hat{\mathbf{r}}_{yx}(n)$ as:

$$\hat{\mathbf{r}}_{yx}(n) = (1-\alpha_x)\hat{\mathbf{r}}_{yx}(n-1) + \alpha_x\mathbf{y}(n)\hat{X}^*_{ref}(n) \tag{3-9}$$

Where α_x is the smoothing factor for target speech signal, and $\hat{X}_{ref} = G(X_{ref} + V_r)$ indicates the clean speech estimate from the reference microphone. This estimate can be done with any of the existent speech enhancement methods, but it is recommended to use one in which the speech distortion can be controlled and set it as small as possible. In this case it will be used the one found in [17]. This method works as follows:

Sigmoid function with *a priori* SNR

In order to find $\hat{X}_{ref}$ it is needed to chose one of the channels as a reference, in our case the input of the closest microphone to the source can be chosen. This signal must be filtered with one method of speech enhancement.
In [17] it is proposed to use a gain function which can be easily controlled and that has a flexible shape. In this case we use the modified sigmoid gain function:

$$G_{MSIG}(k,n) = \frac{1 - exp[-a_1\hat{\xi}(k,n)]}{1 + exp[-a_1\hat{\xi}(k,n)]} * \frac{1}{1 + exp(-a_2[\hat{\xi}(k,n) - c])} \tag{3-10}$$

Where ξ is the *a priori* SNR defined as

$$\xi(k,n) = \frac{\Phi_x(k,n)}{\Phi_v(k,n)}$$

Which can be estimated as a function of the *a posteriori* SNR and an instantaneous *a priori* SNR, the *a posteriori* SNR and applying a recursive smoothing procedure:

$$\xi_{inst}(k,n) = max\{\gamma(k,n) - 1, 0\}$$

using the fact that $X(k,n-1) = G_{MSIG}(k,n)Y(k,n-1)$:

$$\hat{\xi}(k,n) = \alpha_{SNR}G^2_{MSIG}(k,n-1)\gamma(k,n-1) + (1 - \alpha_{SNR})\xi_{inst}(k,n)$$

The variance in the *a priori* SNR estimate can lead to audible musical noise due to the higher sensitivity to changes. In order to reduce such musical noise, a first order recursive smoothing procedure, as shown in [19], can be applied in the *a posteriori* SNR estimation:

$$\bar{\gamma}(k,n) = \frac{\Phi_y(k,n)}{\hat{\Phi}_v(k,n)}$$

With:

$$\Phi_y(k,n) = \alpha_{y'}\Phi_y(k,n-1) + (1 - \alpha_{y'})|Y(k,n)|^2$$

The resulting function of equation (3-10) can have different shapes as shown in **3-2** according to the choice of the constants a_1, a_2 and c, for this implementation the values were chosen as function MSIG-fix3:

$$a_1 = 15$$

$$a_2 = 0.6351$$

$$c = 0.2243$$

Functions	Parameters			Fitted curve
	a_1	a_2	c	
MSIG-fix1	2.3918	0.2120	−1.7071	LSA
MSIG-fix2	11.6869	0.4337	0.7556	WF
MSIG-fix3	15	0.6351	0.2243	SIG

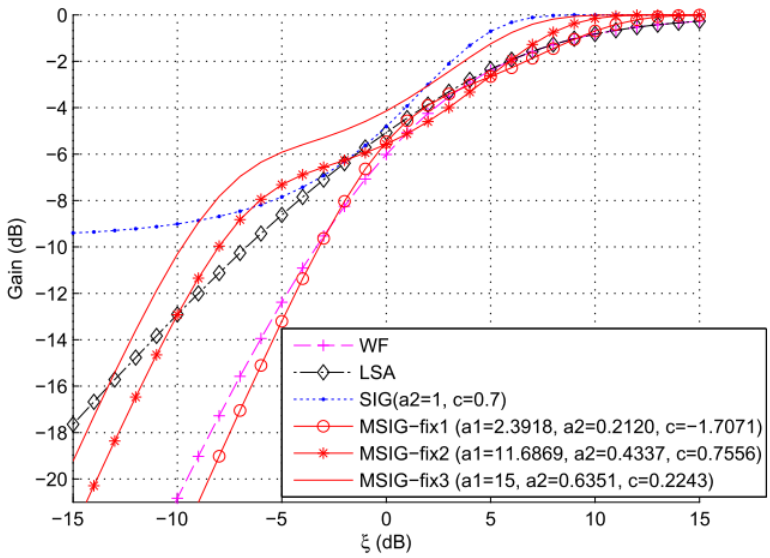

Figure 3-2: Different shapes of the gain function, taken from [17].

3.3 Speech Presence Probability

Several authors have made important research on SPP methods.
All real time audio processing algorithms are implemented with short time Fourier transformations (STFT). The speech presence probability method, has as output a vector of the same size than each frame vector, each element of this vector is a real number with values from 0 to 1 which denote the probability of finding voice in each frequency bin. In this case 0 is the lowest probability of finding voice and 1 is the highest. In the figure **3-3** it is shown the spectrogram of a clean female voice signal and in figure **3-4** it is shown all the SPP vectors of this signal.
It is possible to see that, in an ideal SPP (absence of noise), the output has the value of 1 in each bin where there is presence of speech in the signal. Note that in an ideal case all speech presence values should be 1 and this does not depend on the magnitude of the voice energy in the signal, following the same logic, in silent segments or nose segments, the values should be 0.
In this project, different algorithms for finding the SPP will be tested under different SNR conditions. For testing this, the SNR of the output of the MWF will be compared for each SPP algorithm.

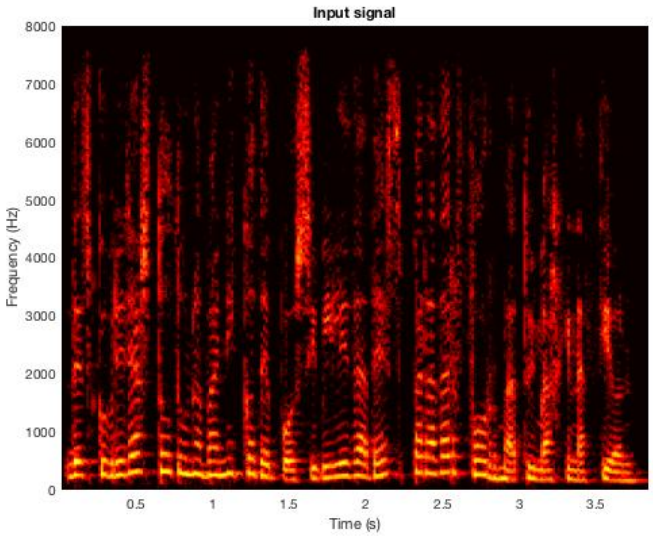

Figure 3-3: Spectrogram of a clean voice audio file.

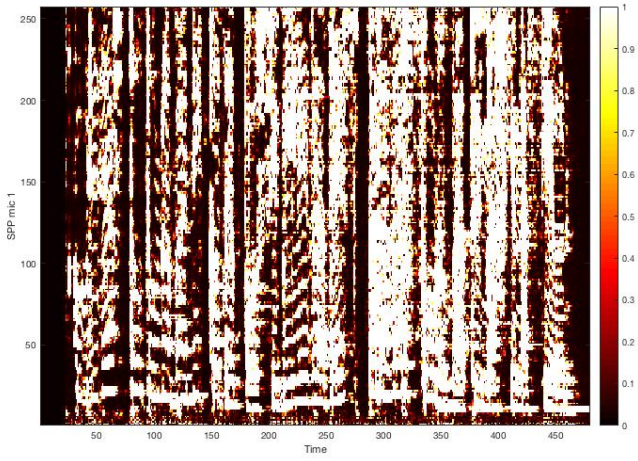

Figure 3-4: Speech presence probability of a clean voice audio file.

4 Speech Presence Probability Implementation

In this chapter it will be shown how the SPP method found in [5] can be implemented in Matlab so that the different variations in the algorithm can be tested later.

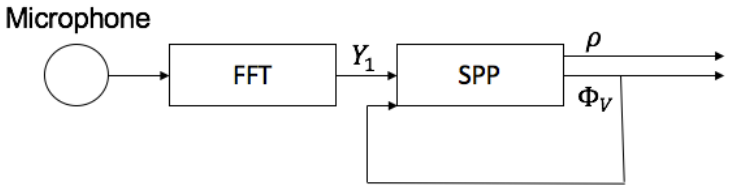

Figure 4-1: Block diagram of SPP method.

4.1 Implementation

Under the assumption that a soft decision VAD might be more precise than the hard decision, the following algorithm will be included in the multi-channel noise reduction method. Following this, the SPP was found as proposed in [5]:

$$\rho = \frac{1}{\left(1 + (1+\xi_{opt}) exp\left(-\frac{|Y_1(k,n)|^2}{\Phi_V(k,n-1)} \frac{\xi_{opt}}{\xi_{opt}+1}\right)\right)} \tag{4-1}$$

Where ρ is the variable that takes values between 0 and 1 indicating the probability of voice presence (0 means no voice, 1 means total certitude of voice) and ξ_{opt} is a fixed a priori SNR, Y_1 is the input , Φ_V is the noise power density, and:

$$\Phi_{V,SPP}(k,n) = \rho\Phi_{V,SPP}(k,n-1) + (1-\rho)|Y_1(k,n)|^2 \tag{4-2}$$

After this, the value of $\Phi_{N,SPP}$ is integrated to the estimated noise power density:

$$\Phi_V(k,n) = \alpha_{N,SPP}\Phi_V(k,n-1) + (1-\alpha_{N,SPP})\Phi_{V,SPP}(k,n) \tag{4-3}$$

In [11] it is proposed a similar method in which the value of ρ is modeled as a sigmoid function, with this, it is possible to adjust its slope (a_{sig}) and mean (c_{sig}). The value of ρ wil be given as:

$$\rho = \frac{1}{\left(1 + exp\left(-a_{sig}\left(\frac{|Y_1(k,n)|^2}{\Phi_V(k,n-1)} - c_{sig}\right)\right)\right)} \tag{4-4}$$

where:

$$a_{sig} = \frac{\xi_{opt}}{\xi_{opt} + 1}$$
$$c_{sig} = log\left(\frac{P(\mathcal{H}_0)}{P(\mathcal{H}_1}(1 - \xi_{opt})\right)\frac{\xi_{opt} + 1}{\xi_{opt}}$$

In this project both SPP methods will be tested and compared when they are used with as part of the MWF noise reduction method.

4.2 Stagnation

This algorithms give a very close estimation of the ideal SPP, but, it is susceptible to have stagnation problems. It can be seen that if the spectral noise power is underestimated, in equation (4-1), it may lead to $\rho = 1$ even if the power of the input is low with respect to the real and unknown noise power. In this case the equation (4-2) won't update the noise power anymore and the noise will remain underestimated. This is know as stagnation.
Both described methods also propose a way to avoid stagnation which will be explained in this chapter and part of the testing algorithms.

Tracking Time Average

The algorithm showed in [5] propose to force a lower the value of ρ when it's time average is higher than certain threshold. This means, if ρ has high values for too long it's forced to a lower one, this can be expressed as:

$$\rho = \begin{cases} min(0.5, \rho) & \bar{\rho} > 0.9 \\ \rho & else \end{cases} \tag{4-5}$$

3 Regions

In [11], the SPP was categorized in three regions which control the update speed of the noise PSD estimate from equation 4-3 as follows:

$$\tau = \begin{cases} \mathcal{P}_1 & if \quad \rho \leq 0.3 \\ \mathcal{P}_2 & if \quad 0.3 < \rho < 0.6 \\ min(\mathcal{P}_3, \rho) & if \quad \rho \geq 0.6 \end{cases} \tag{4-6}$$

After, finding the right τ, the smoothing value α is found with the following equation:

$$C(\tau_m) = \alpha = e^{\frac{-2.2R}{F_s * \tau}} \tag{4-7}$$

The proposed time values for $\mathcal{P}_i$ are:

$$\mathcal{P}_1 = 50ms$$

$$\mathcal{P}_2 = 80ms$$

$$\mathcal{P}_3 = 240ms$$

5 Evaluation

In this chapter it is shown the steps for implementing the MFW, the evaluation score that will be use for testing it and the results of testing it under different conditions.

5.1 Algorithm

The MWF noise reduction method was implemented in Matlab following these steps:

1. Compute the auto and cross PSD of both inputs.
2. Find the single microphone SPP ρ and noise PSD with (4-1) or (4-4) and (4-2)
3. Apply a single channel noise reduction method to the reference.
4. Obtain correlation matrices with (3-6) and (3-7).
5. Obtain the estimated matrix $\hat{\mathbf{r}}_{yx}$ with (3-9) using the filtered reference of step 3.
6. Obtain filter $\mathbf{W}_{MWF_\lambda}$ with (3-8).
7. Obtain Z as $Z(k,n) = \mathbf{W}^H_{MWF_\lambda}(k,n)\mathbf{y}(k,n)$.

The MWF was implemented in Matlab with both previously showed SPP algorithms and with both methods to avoid stagnation. Here it will be shown how to evaluate the performance of this implementation and the results of it.

5.2 Data Base

To test in the most accurately possible way all the noise reduction methods, it is necessary to have a proper data base and a proper algorithm to run the measurement metric. For this is was used a radio studio with 4 microphones, one main speaker and a set of several speakers.
The data base consists of 10 voices (5 female and 5 male) and 5 different kind of usual noises. The main speaker and microphones were arranged as shown in figure **5-1** where $D = 50cm$ and $d = 10cm$. The main speaker was reproducing the clean voice files and the set of speakers was placed all around the studio to create a good simulation of spatially homogeneous distributed noise.

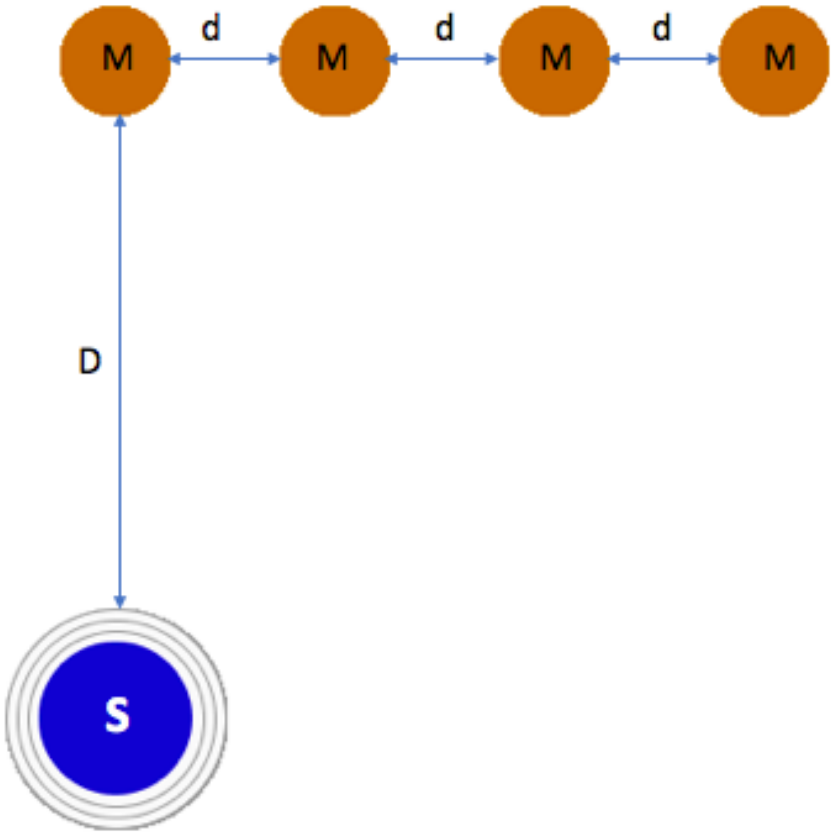

Figure 5-1: Microphone array and audio source positions.

5.3 Perceptual Evaluation of Speech Quality

As shown in [15], the measure of PESQ (Perceptual Evaluation of Speech Quality) is a measure which evaluates the similarity between two speech signals in a perceptual scale. This means that a high value of PESQ means a good speech enhancement method.

The score of PESQ is given in dBs, where the higher score means better speech enhancement.

The Matlab code capable of performing the PESQ evaluation between two signals was taken from [7].

In order to obtain accurate scores, two files from the voice data base were taken (one female voice and one male voice) and two from the noise database and each voice was mixed with each noise in a range of SNR of $[-30dB, 15dB]$. These signals were taken as imputs for the different filters, and then the output of each filter was plotted and compared with the other methods and with the original noisy signal.

5.4 Results

The results of the evaluation process are shown in figures **5-2**, **5-3**, **5-4**, **5-5**. The method labeled as "Gerk" refers to the one found in [5] and the one labeled as "Yong" refers to the one found in [12] and the label "Noisy" is the noisy input with no filter. The use of these methods with each of the algorithms to avoid stagnation and their combination were plotted and their PESQ score is displayed.

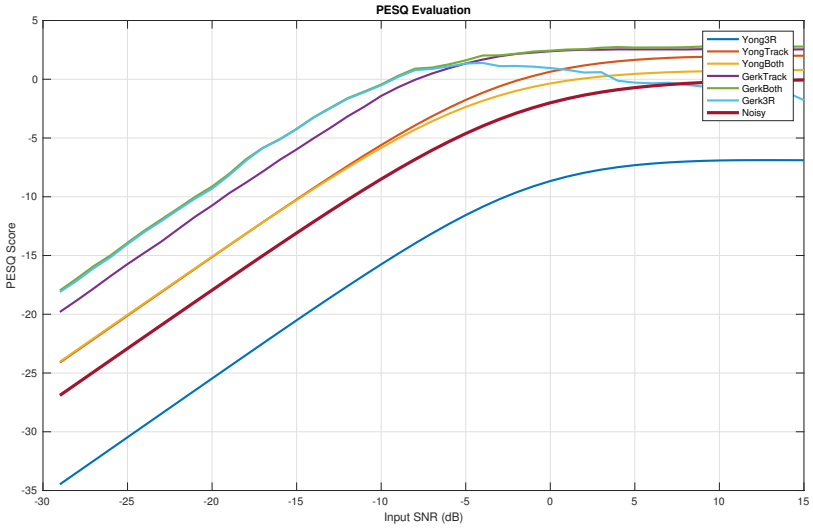

Figure 5-2: PESQ results for female voice and car noise audios.

It can be seen in all figures that the combination of "Yong" with the 3 regions method leads to an important voice degradation and its PESQ values are always lower than the noisy input, however when this method is used with the tracking time algorithm, the score increases and there is speech enhancement. This can be caused by the speed in which this method updates the SPP value not being compatible with the 3 regions. In this case, combining both algorithms for avoiding stagnation doesn't improve the value.

It can also be seen that when the method "Gerk" is combined with the 3 regions algorithm, the performance for low SNR values is always enhancing the speech but when the SNR gets higher, the scores decreases because of speech degradation. The tracking time method can help to fix this, it is shown that this method has high PESQ values for high SNR input values also when mixed with the 3 regions method.

In general, the method shown in [5] shows a better performance when being implemented with the MWF. Also it can be noticed that the 3 regions method tends to have better performance for

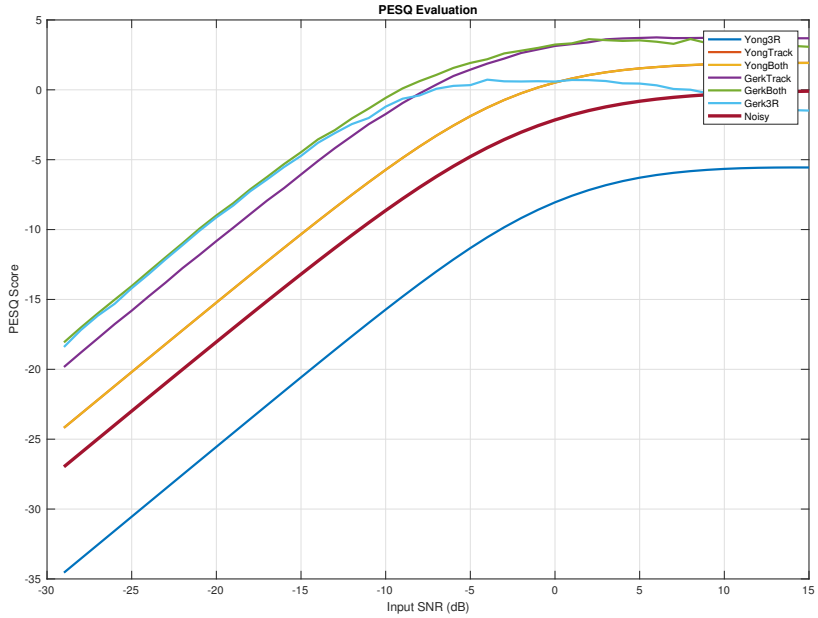

Figure 5-3: PESQ results for male voice and car noise audios.

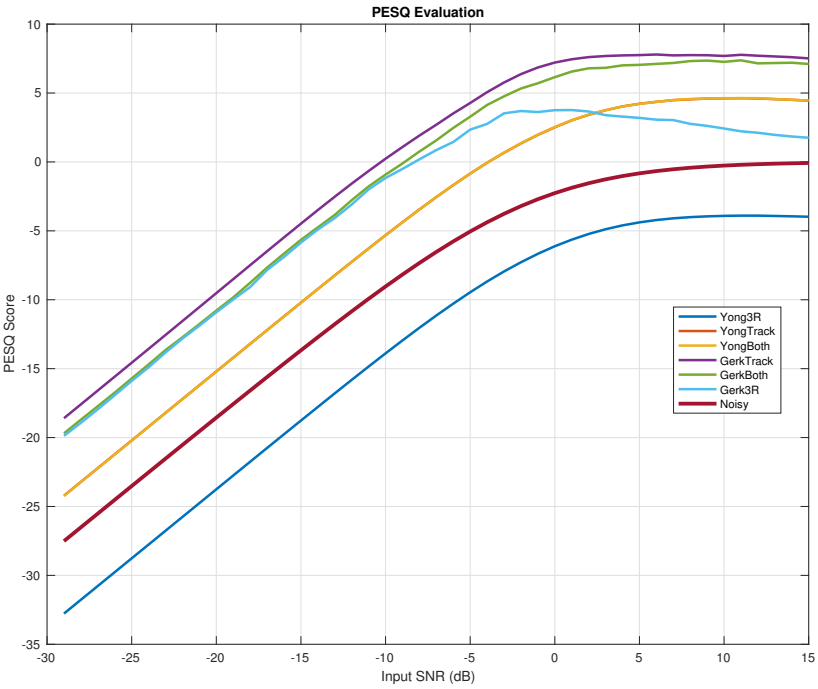

Figure 5-4: PESQ results for female voice and street noise audios.

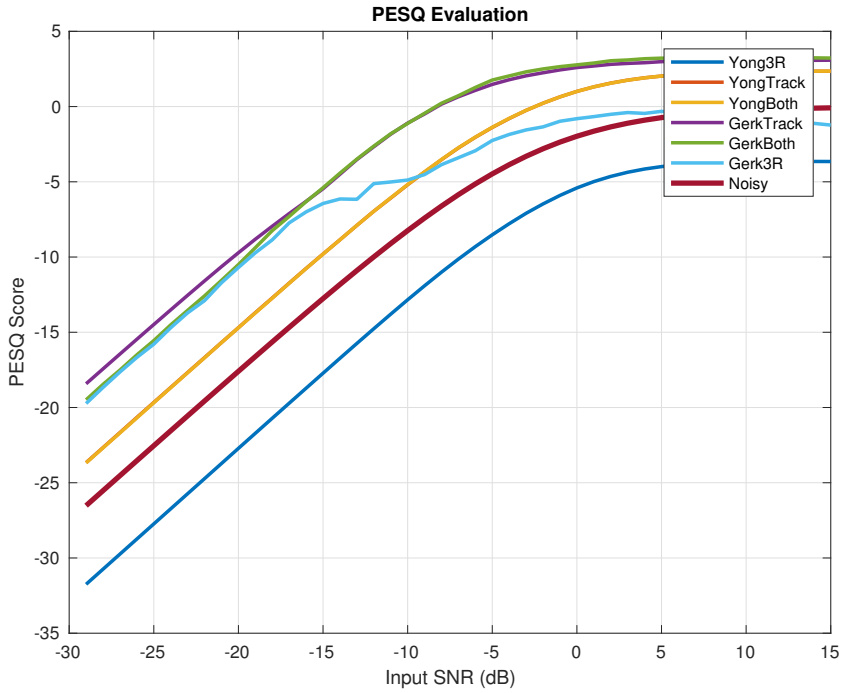

Figure 5-5: PESQ results for male voice and street noise audios.

low SNR input values over the tracking time method but this is reversed for higher SNR values. In general, mixing these two algorithms for avoiding stagnation leads to good speech enhancement results. It is just important to notice that the tracking method will have more memory storage.

In the figure **5-6** is shown the input and output spectogram of one of the test. There it is notorious how big part of the noise is attenuated while preserving the voice. In figure **5-7** the SPP values of this case are show, it is notorious that there is voice over-estimation (noise taken as voice), but not under-estimation (most of the voice is detected), which is one of the conditions for having a correct reference for the multichannel filter.

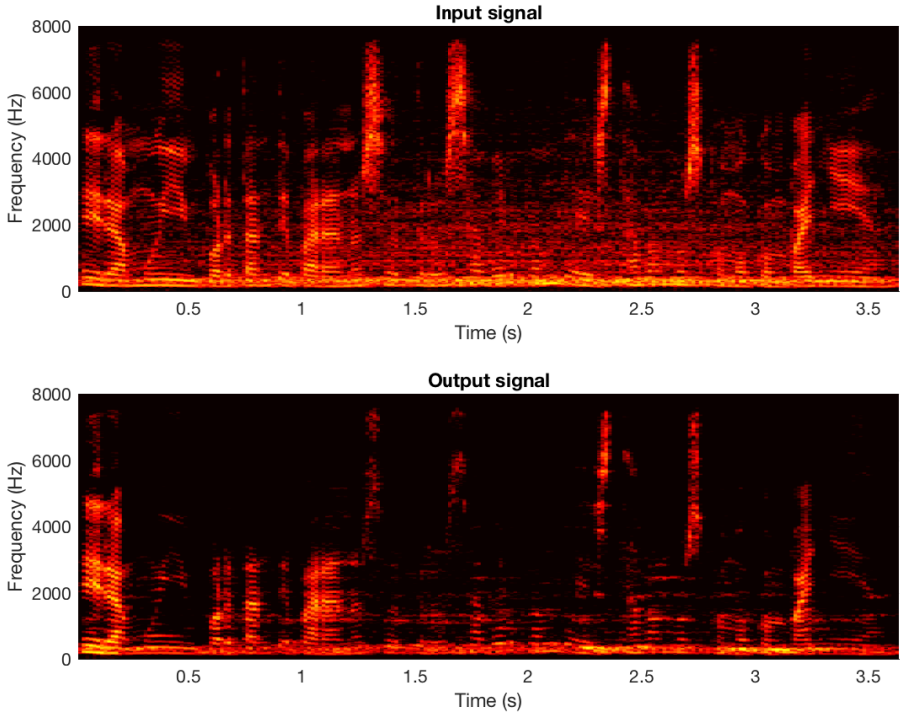

Figure 5-6: Input and out spectrograms when using the MWF with ”Gerk” method and both algorithms for avoiding stagnation.

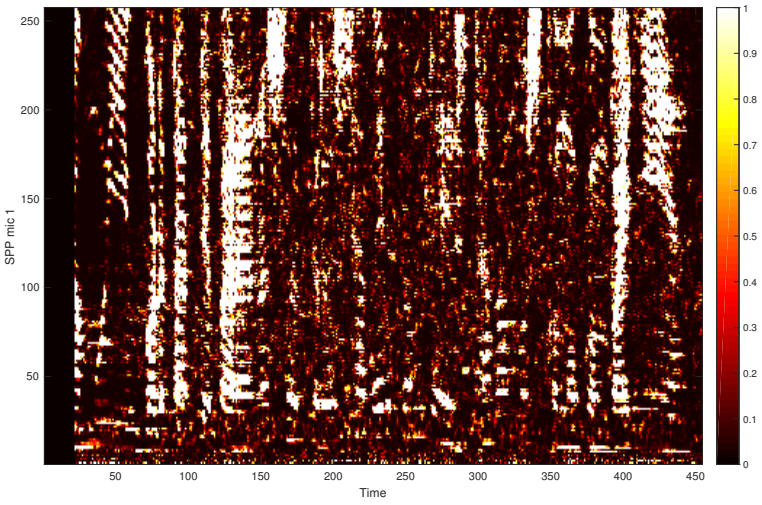

Figure 5-7: SPP values.

6 Conclusions and Recommendations

6.1 Conclusions

In this document it was presented a detailed illustration of the Multichannel Wiener Filter as well as a description of each of it's blocks and how to implement them. Special attention was given to the Speech Presence Probability block and different algorithms for it were tested. The overall result shows that the method proposed in [5] when tracking time average and with the 3 regions method for avoiding stagnation.

The method proposed in [12] showed that is not compatible with the 3 regions method and the result of this was a lower PESQ score than the noisy input, which means that was only degrading the voice. However, when used with the tracking time average method, the results were better, but still worst that [5].

The only consideration that may be done at the moment of choosing an algorithm for avoiding stagnation for [5] is that the time tracking average needs more storage memory due to it's need to save the complete vector of the smoothed SPP for every frame, which could lead to higher costs depending the implementation.

6.2 Recommendations

This complete document can be used a guide for future projects, here it can be found step by step how to implement a multichannel noise filter, a single channel noise filter, two speech presence probability methods with two different algorithms for avoiding stagnation and references on how to test any speech enhancement method. All of this can save important time for further research.

Also, because of the big amount of blocks that this method has, it is always possible to try to find optimum values for different situations. In the future, values like the a priori SNR and smoothing times can be tested and arranged for more specific situations.

These kind of algorithms is be important for communications systems and this document can be used as help for coding speech enhancement projects in mobile phones, car hands-free systems,

computers, microphone arrays for public speeches and many others.

Finally, the data base and the code are open to be used for tests.

7 Bibliography

[1] Abdelaziz, Trabelsi ; Franċois-raymond, Boyer ; Yvon, Savaria: Real-Time Dual-Microphone Speech Enhancement. (2005)

[2] Bourgeois, Julien ; Linhard, Klaus: Frequency-domain multichannel signal enhancement: minimum variance vs. minimum correlation. In: *Signal Processing Conference, 2004 12th European* (2004)

[3] Ephraim, Yariv ; Malah, David: Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 32 (1984), Nr. 6, S. 1109–1121. http://dx.doi.org/10.1109/TASSP.1984.1164453. – DOI 10.1109/TASSP.1984.1164453

[4] Faneuff, Jeffery J.: Spatial, Spectral, and Perceptual Nonlinear Noise Reduction for Hands-free Microphones in a Car. (2002), Nr. July

[5] Gerkmann, Timo ; Hendriks, Richard C.: Noise power estimation based on the probability of speech presence. In: *IEEE Workshop on Applications of Signal Procesing to Audio and Acoustics* (2011), S. 145–148. http://dx.doi.org/10.1109/ASPAA.2011.6082266. – DOI 10.1109/ASPAA.2011.6082266

[6] Joer Meyer, Klaus Uwe S.: *Meyer, Simmer - 1997 - Multichannel speech enhancement in a car environment using wiener filtering and spectral subtraction.pdf*

[7] Loizou, Philipos C.: *Speech Enhancement, theory and practice*. Bd. Second edi

[8] Luo, H.Y. ; Denbigh, P.N.: A speech separation system that is robust to reverberation. In: *Proceedings of ICSIPNN '94. International Conference on Speech, Image Processing and Neural Networks* (1994), Nr. April, S. 339–342. http://dx.doi.org/10.1109/SIPNN.1994.344897. – DOI 10.1109/SIPNN.1994.344897. ISBN 0–7803–1865–X

[9] Mei, Tiemin ; Xi, Jiangtao ; Yin, Fuliang ; Mertins, Alfred ; Chicharo, Joe F.: Blind source separation based on time-domain optimization of a frequency-domain independence criterion. In: *IEEE Transactions on Audio, Speech and Language Processing* 14 (2006), Nr. 6, S. 2075–2085. http://dx.doi.org/10.1109/TASL.2006.872623. – DOI 10.1109/TASL.2006.872623

[10] Nelke, Christoph M. ; Beaugeant, Christophe ; Vary, Peter: Dual microphone noise PSD estimation for mobile phones in hands-free position exploiting the coherence and speech presence probability. In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* (2013), Nr. May, S. 7279–7283. http://dx.doi.org/10.1109/ICASSP.2013.6639076. – DOI 10.1109/ICASSP.2013.6639076

[11] Pei Chee Yong, Hai Huyen Dam C. Sven Nordholm N. Sven Nordholm: Noise Estimation Based on Soft Decisions and Conditional Smoothong. (2012), Nr. September, S. 4–6

[12] Pei Chee Yong, Sven Nordholm, Hai Huyen Dam C.: Noise Estimation Based on Soft Decisions and Conditional Smoothong. (2012), Nr. September, S. 4–6

[13] Rosca, J ; Balan, R ; Fan, Np ; Beaugeant, C ; Gilg, V: Multichannel voice detection in adverse environments. In: *of EUSIPCO 2002* 1 (2002), Nr. 2, 2–5. http://www.cs.rochester.edu/u/www/u/rosca/preprints/2002/eusipco2002mvd.pdf

[14] Rubio, Juan E. ; Ishizuka, Kentaro ; Sawada, Hiroshi ; Araki, Shoko ; Nakatani, Tomohiro ; Fujimoto, Masakiyo: Two-microphone Voice Activity Detection based on the homogeneity of the direction of arrival estimates. In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* 4 (2007), Nr. November 2015. http://dx.doi.org/10.1109/ICASSP.2007.366930. – DOI 10.1109/ICASSP.2007.366930

[15] Sharma, Dushyant ; Meredith, Lisa ; Lainez, Jose ; Barreda, Daniel ; Naylor, Patrick A.: A non-intrusive PESQ measure. In: *2014 IEEE Global Conference on Signal and Information Processing, GlobalSIP 2014* (2014), S. 975–978. http://dx.doi.org/10.1109/GlobalSIP.2014.7032266. – DOI 10.1109/GlobalSIP.2014.7032266. ISBN 9781479970889

[16] Smaragdis, Paris: Blind separation of convolved mixtures in the frequency domain. In: *Neurocomputing* 22 (1998), Nr. 1-3, S. 21–34. http://dx.doi.org/10.1016/S0925-2312(98)00047-2. – DOI 10.1016/S0925–2312(98)00047–2. – ISBN 0925–2312

[17] Yong, Pei C. ; Nordholm, Sven ; Dam, Hai H.: Optimization and evaluation of sigmoid function with a priori SNR estimate for real-time speech enhancement. In: *Speech Communication* 55 (2013), Nr. 2, 358–376. http://dx.doi.org/10.1016/j.specom.2012.09.004. – DOI 10.1016/j.specom.2012.09.004. – ISSN 01676393

[18] Yong, Pei C. ; Nordholm, Sven ; Dam, Hai H. ; Leung, Yee H. ; Lai, Chiong C.: Incorporating multi-channel Wiener filter with single-channel speech enhancement algorithm. In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* (2013), S. 7284–7288. http://dx.doi.org/10.1109/ICASSP.2013.6639077. – DOI 10.1109/ICASSP.2013.6639077. – ISBN 9781479903566

[19] Yong, Pei C. ; Nordholm, Sven ; Dam, Hai H. ; Low, Siow Y.: On the optimization of sigmoid function for speech enhancement. In: *European Signal Processing Conference* (2011), Nr. Eusipco, S. 211–215. – ISSN 22195491